LSTAT2170 - Time Series

Modeling a time series for the Monthly temperature in Recife (Brazil)

Rousseau Mathieu, 67001800

UCLouvain
Belgium
12/05/2023

1 Introduction

The aim of this project is to analyse the mean monthly temperature in (°) in the city of Recife (Brazil) from 1986 to 1995.

First, we will begin by an analysis of the data. We will first check for any trend, seasonalities or variability of the variance and apply ad-hoc method in order to be able to analyze the ACF and PACF plots of the data. These plots should give us a first intuition of a possible model to model our data. We will then use automatic model selection using different criterions like the AIC, BIC in order to confirm our intuition and help us to select one or two final models. We will then test our model(s) ...

Finally, we will ensure our model has good prediction power use it to predict the temperature for the city of Recife in 1996.

2 Basic analysis of the data

2.1 Analyse of trend and seasonalities

Looking at the plot of the time serie for the mean monthly air temperature in Recife below (**??**), we directly notice a seasonality in the data. This was obviously expected as the temperatures fluctuate between seasons. However, this is not clear if there is any trend (*there seems to be a decrease in the mean monthly temperature which is kind of unexpected considering the climate change taking place since the beginning of the industrial era*) nor an instability in the variance. In order to have a clearer view of these elements, we can decompose the time serie by performing a **classical decomposition**. This method assume that the seasonal component repeat from year to year which is kind of a reasonable assumption for air temperature data. Let $\{ X_t, t \in \mathbb{Z} \}$ be our time serie, assuming an additive decomposition we can write,

$$X_t = S_t + T_t + R_t \tag{1}$$

with S_t, T_t and R_t respectively the seasonal, the trend-cycle and the remainder components.

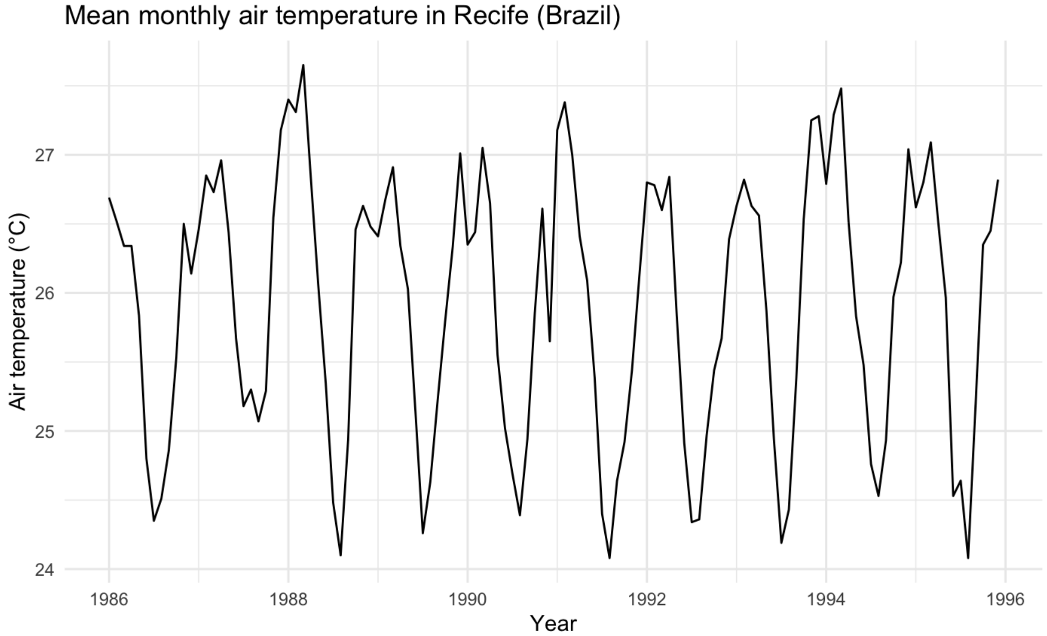


FIGURE 1 – *Time series for the mean monthly air temperature in the city of Recife (Brazil) for the years* 1986 – 1995

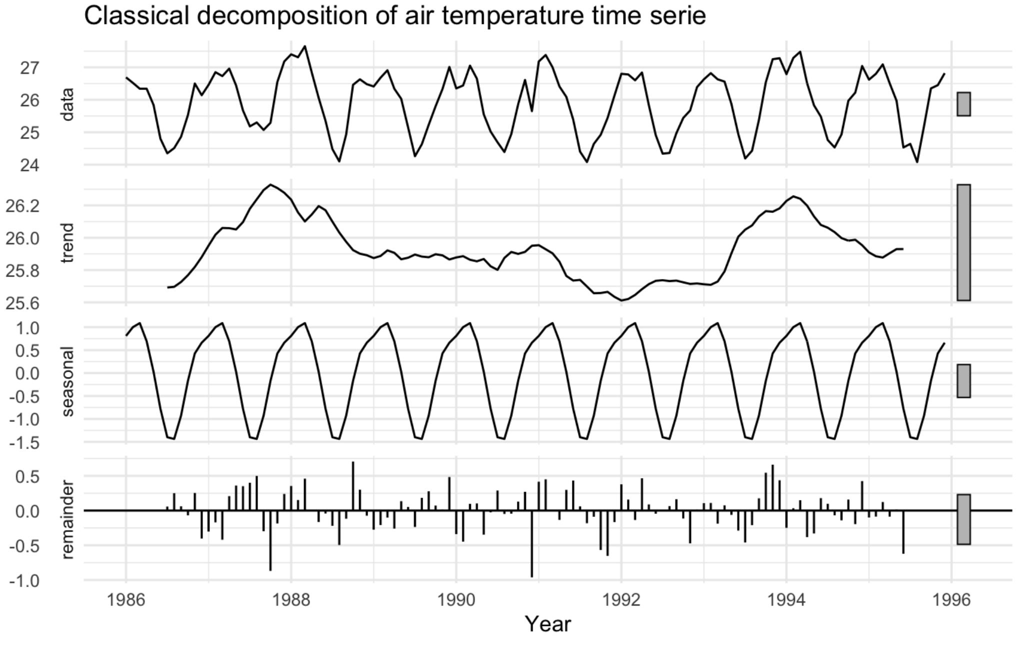


FIGURE 2 – *Decomposition of the time serie X_t into seasonal, trend-cycle and reminder components*

Being situated in the southern hemisphere, the Brazil has its seasons inverted compared to our regions. This country is situated near the equator and therefore has pretty high temperature all the year long. However, the highest temperatures happen between the end and the beginning of the year and the lowest toward the middle of the year. That's clearly what we can notice in the seasonal part (**??**). We can also notice a decreasing trend in the temperatures between 1988 and 1993 before it rises up again.

An interesting thing we can do is to emphasis the seasons with a seasonal plot. We notice in particular that the seasonal variations are roughly constants over the years. That's why we can assume an additive decomposition.

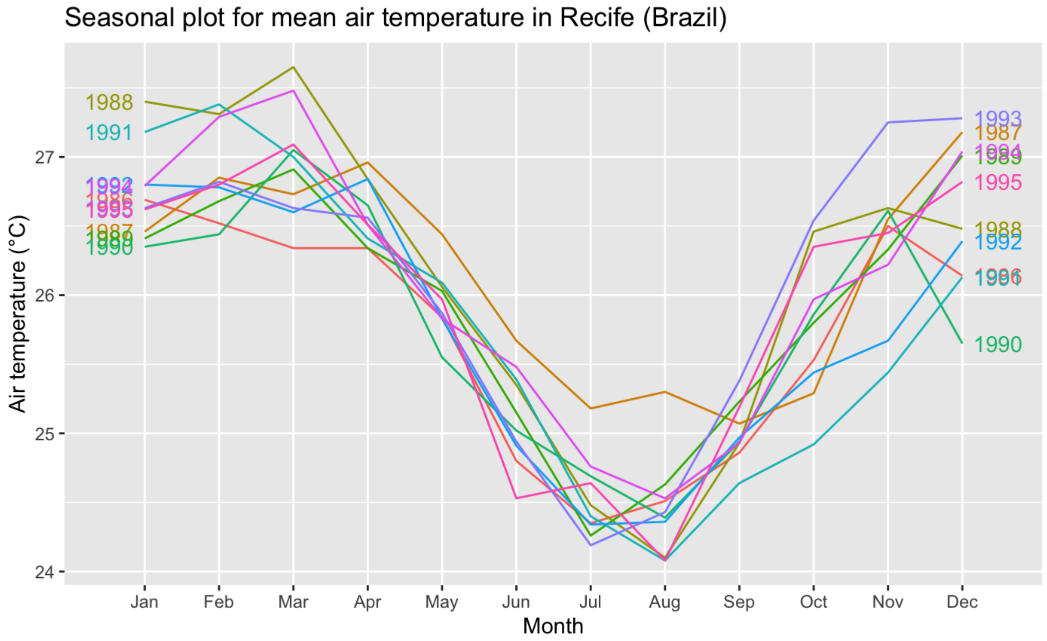


Figure 3 – *Seasonal plot for the mean air temperature in Recife (Brazil)*

3 Box-Jenkins analysis

Before going further and select a model that will be able to predict the air temperature in Recife for the months following the year 1995, we need to detrend and deseasonalize our time serie.

3.1 Differencing

We first remove the seasonality by taking a seasonal difference which consist by taking the difference between an observation and the previous observation at the same season (in our case, 12 months before) : $X'_t = X_t - X_{t-12}$. Then, we remove the trend by first differencing the serie.

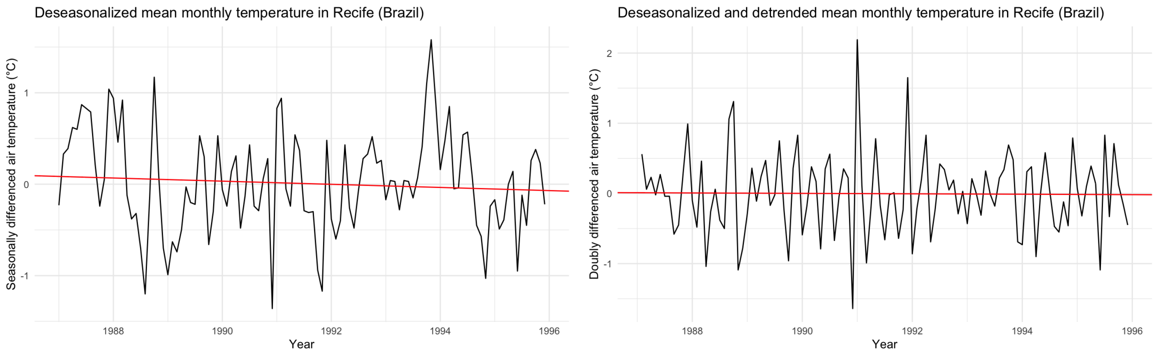


Figure 4 – *Deseasonalized time serie (at left) and doubly differenced time serie (deseasoned and detrended) (at right). In red is the regression line.*

After doubly differencing, the time serie is now stationary as shown on the plots above.

3.2 Model intuition with ACF and PACF plots

We can now go further and analyse the autocorrelation and partial autocorrelation plots (ACF, PACF) of our stationary time series to have an intuition about the possible model we could choose to modelize our serie.

Because we are dealing we seasonal data, we will probably use a SARIMA model which consists in adding additional seasonal terms to the ARIMA model,

$$\underbrace{(p, d, q)}_{\text{non-seasonal part}} \times \underbrace{(P, D, Q)_m}_{\text{seasonal part}} \tag{2}$$

where **p**, **d** and **q** (**P**, **D** and **Q**) are respectively the orders of the non-seasonal (seasonal) AR-process, difference and MA-process whereas **m** is the number of observations per year (in our case $m = 12$).

On a yearly basis, up to a lag of 8 years, the ACF converge toward zero likewise the PACF. However, we notice a significant spike at lag 12 for the ACF suggesting a seasonal MA(1) component. In the same idea, the PACF shows a significant spike at lag 12 as well suggesting a seasonal AR(1) component.

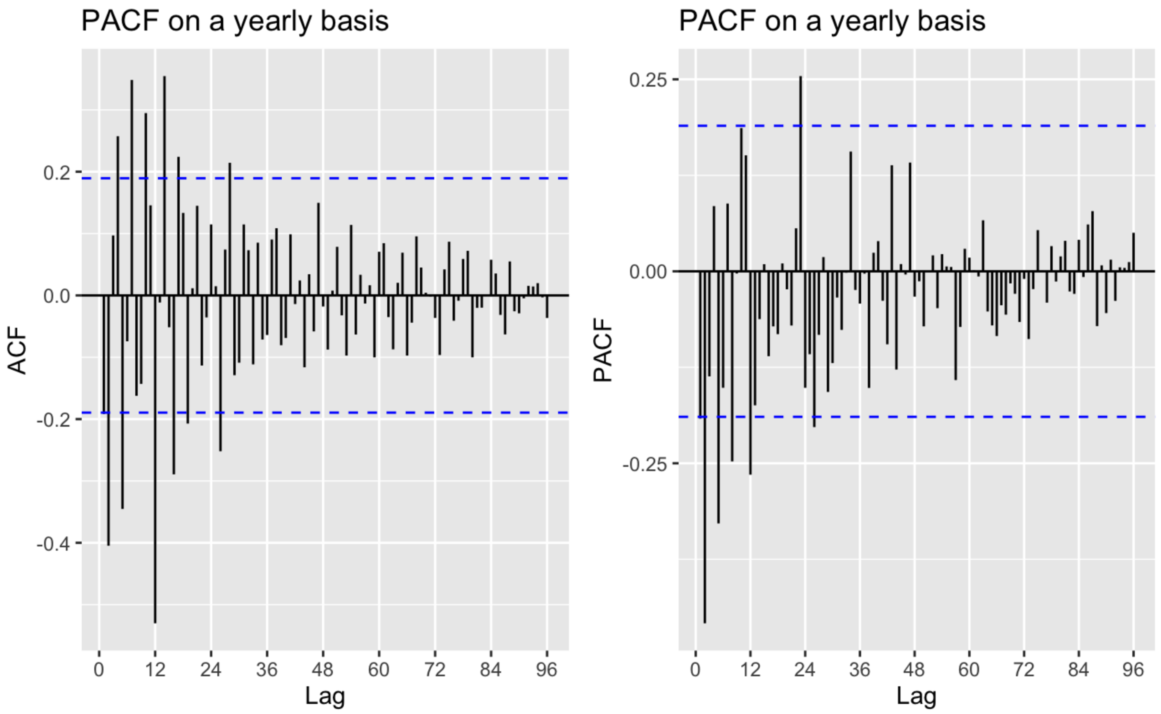


Figure 5 – *ACF and PACF plots on a yearly basis*

We can restreint ourselve to a monthly basis to check for any non-seasonal components. We have 5 significant spikes respectively at lag 2, 4, 5, 7 and 10 for the ACF suggesting a non-seasonal MA(5) component. On the other hand, we have 3 significant spikes at lag 2, 5 and 8 for the PACF suggesting a non-seasonal AR(3) component.

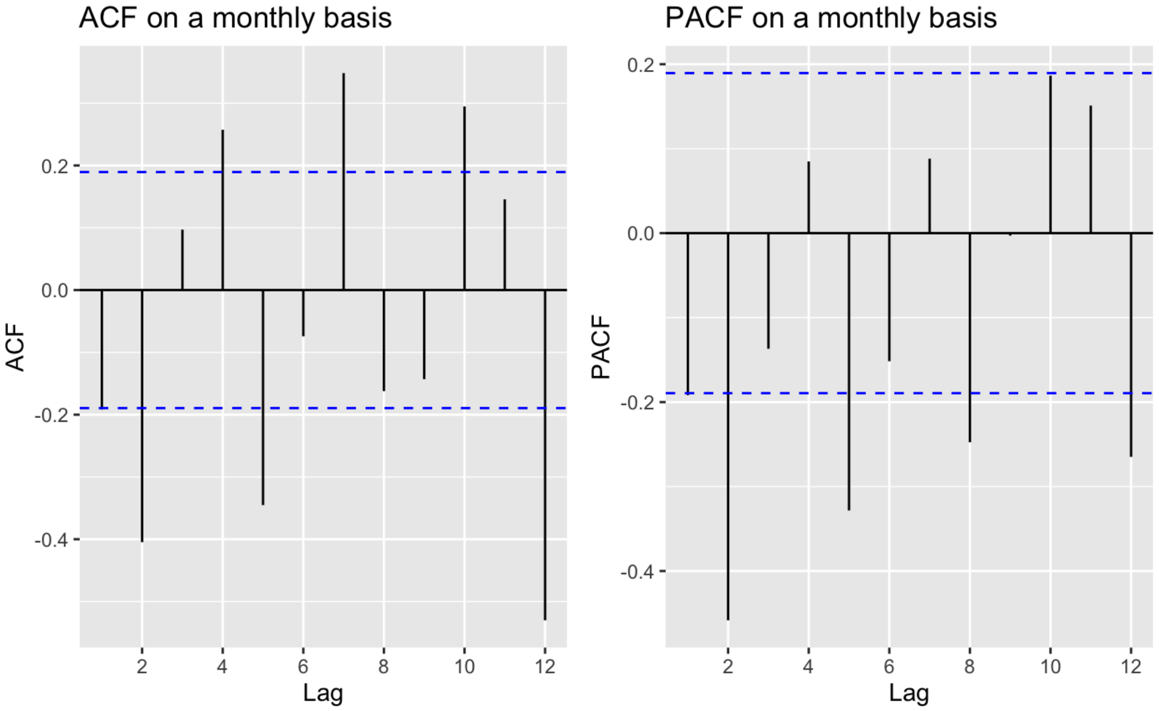


Figure 6 – *ACF and PACF plots on a monthly basis*

Finally, with all these considerations, our first intuition is to modelize our time serie with the following SARIMA process,

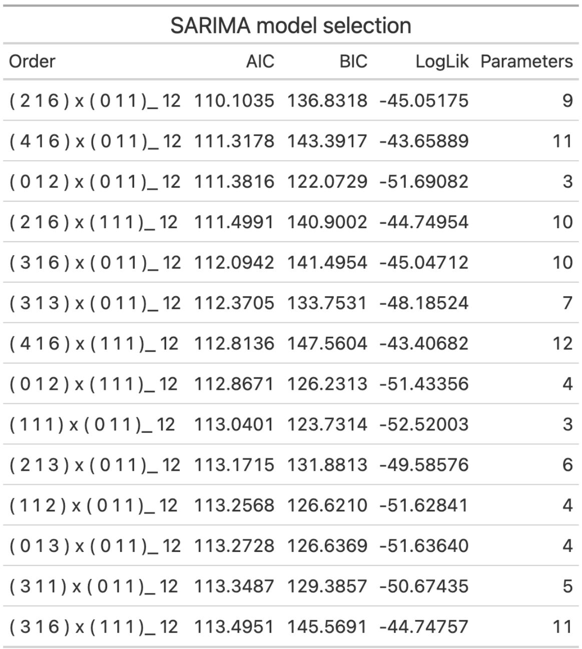
$$(3,1,5) \times (1,1,1)_{12} \tag{3}$$

But because of the decay in ACF and PACF on a yearly basis, we could as well modelize with these two other SARIMA processes,

$$(3,1,5) \times (0,1,1)_{12} \quad ; \quad (3,1,5) \times (1,1,0)_{12} \tag{4}$$

3.3 Automatic model selection with AIC and BIC

After guessing the ideal model, we perform model selection for p and q up to 4 and 6 respectively because both ACF and PACF plots have at least one spike near outside the confidence interval. We also choose P and Q both up to 1. Below are the top 10% models based on the AIC criterion. We chose to also compute the BIC (as the AIC tend to overestimate the number of parameters) and the log likelihood.



SARIMA model selection

Order	AIC	BIC	LogLik	Parameters
(2 1 6) x (0 1 1)_ 12	110.1035	136.8318	-45.05175	9
(4 1 6) x (0 1 1)_ 12	111.3178	143.3917	-43.65889	11
(0 1 2) x (0 1 1)_ 12	111.3816	122.0729	-51.69082	3
(2 1 6) x (1 1 1)_ 12	111.4991	140.9002	-44.74954	10
(3 1 6) x (0 1 1)_ 12	112.0942	141.4954	-45.04712	10
(3 1 3) x (0 1 1)_ 12	112.3705	133.7531	-48.18524	7
(4 1 6) x (1 1 1)_ 12	112.8136	147.5604	-43.40682	12
(0 1 2) x (1 1 1)_ 12	112.8671	126.2313	-51.43356	4
(1 1 1) x (0 1 1)_ 12	113.0401	123.7314	-52.52003	3
(2 1 3) x (0 1 1)_ 12	113.1715	131.8813	-49.58576	6
(1 1 2) x (0 1 1)_ 12	113.2568	126.6210	-51.62841	4
(0 1 3) x (0 1 1)_ 12	113.2728	126.6369	-51.63640	4
(3 1 1) x (0 1 1)_ 12	113.3487	129.3857	-50.67435	5
(3 1 6) x (1 1 1)_ 12	113.4951	145.5691	-44.74757	11

Figure 7 – *Summary of the 10% best SARIMA models based on the AIC*

We notice that our intuition was not that bad but we will favour simpler models. Therefore, based on these results, we choose the third model on the list because it has way fewer parameters than the two firsts. The BIC for that model is also the lowest in that list. Thus, that is the following SARIMA-process,

$$\text{model 1 :} \quad (0,1,2) \times (0,1,1)_{12} \tag{5}$$

that presents only 3 parameters.

Nevertheless, we will compare it with a model that has one more parameter,

$$\text{model 2 :} \quad (0,1,2) \times (1,1,1)_{12} \tag{6}$$

We could also compare it with,

$$\text{model 3 :} \quad (1,1,1) \times (0,1,1)_{12} \tag{7}$$

3.4 Model validation

3.4.1 Coefficients significance

We first want to verify that the coefficients of our model are statistically significants. To do that, we perform the following univariate two sided hypothesis test which is based on a normal approximation,

$$H_0 : \text{coeff}(i) = 0$$
$$H_1 : \text{coeff}(i) \neq 0$$

for $i = 0, \dots, (p+P) + (q+Q)$.

The results are the summarized in the following tables,

model 1	ma1	ma2	ma3
/	$3.02e-7$	$2.136e-3$	6.35e-6

Table 1 – *p-value for univariate two sided test for coefficients significant of model 1*

model 2	ma1	ma2	sar1	sma1
/	$2.639e-7$	$3.558e-3$	$4.674e-1$	$1.95e-2$

Table 2 – *p-value for univariate two sided test for coefficients significant of model 2*

model 3	ar1	ma1	sma1
/	$7.982e-7$	$8.393e-25$	$2.277e-7$

Table 3 – *p-value for univariate two sided test for coefficients significant of model 3*

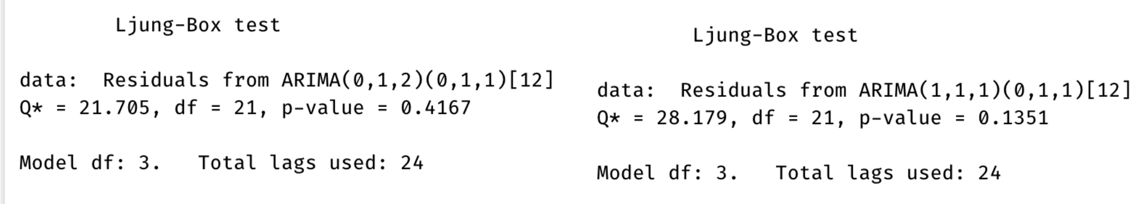
Based on this analysis, for a level of significance of 5%, we can reject our second model as it has one coefficient not statistically significant and we conclude that it is better to consider a model with 3 coefficients.

3.4.2 Residual analysis

We want to check if there is any correlation in the residuals of our models. To do that, we perform a *Ljung-Box test* on the residuals which consists in the following hypothesis,

H_0 : *the residuals are independently distributed.*

H_1 : *the residuals are not independently distributed. They exhibit serial correlation instead.*



```
        Ljung-Box test

data:  Residuals from ARIMA(0,1,2)(0,1,1)[12]
Q* = 21.705, df = 21, p-value = 0.4167

Model df: 3.   Total lags used: 24
```

```
        Ljung-Box test

data:  Residuals from ARIMA(1,1,1)(0,1,1)[12]
Q* = 28.179, df = 21, p-value = 0.1351

Model df: 3.   Total lags used: 24
```

Figure 8 – *Ljung-Box test for model 1 (at left) and model 3 (deseasoned and detrended) (at right).*

All the three models pass the Ljung-Box test at 5% significance level. Then, the residuals are not correlated and we should have an accurate prediction interval. Moreover, the residuals are more or less normals (**??**, **??**)

3.4.3 Predictive power

Despite our models passing the Ljung-Box test, we want to ensure they have good predictive power by comparing the last 20% (i.e. last two years) of the original time serie with a forecast for these same 20% using our models.

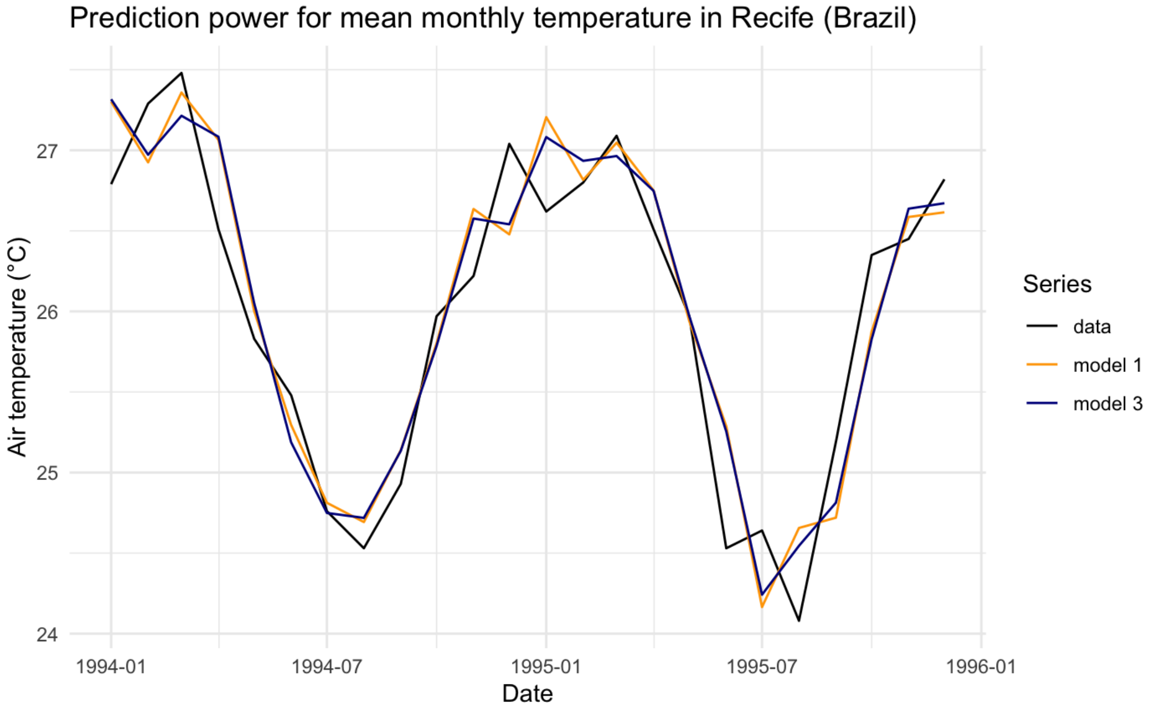
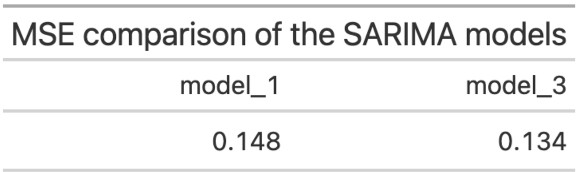


Figure 9 – *Predictive power of the two SARIMA models (1 and 3) compared with the last two years of the original data (in dotted black).*

...

Looking at the MSE of the prediction, we notice that the model 3 is slightly better than model 1 with an MSE of 0.134.



MSE comparison of the SARIMA models

model_1	model_3
0.148	0.134

Figure 10 – *Predictive power of the two SARIMA models (1 and 3).*

Therefore, we choose to keep the model 3 as the final model for forecasting. This model $(1, 1, 1) \times (0, 1, 1)_{12}$ can be written the following way,

$$(1 - \phi_1 B)(1 - B)(1 - B^{12}) = (1 + \theta_1 B)(1 + \Theta_1 B^{12})\varepsilon_t \tag{8}$$

where B is the *"backshift operator"* and $\phi_1 = 0.3529$, $\theta_1 = -0.8946$, $\Theta_1 = -0.9999$.

4 Forecasting

In the last part of this project, we want to forecast the monthly mean air temperature in Recife for the year 1996. We first forecast for the next 12 months using our SARIMA chosen model.

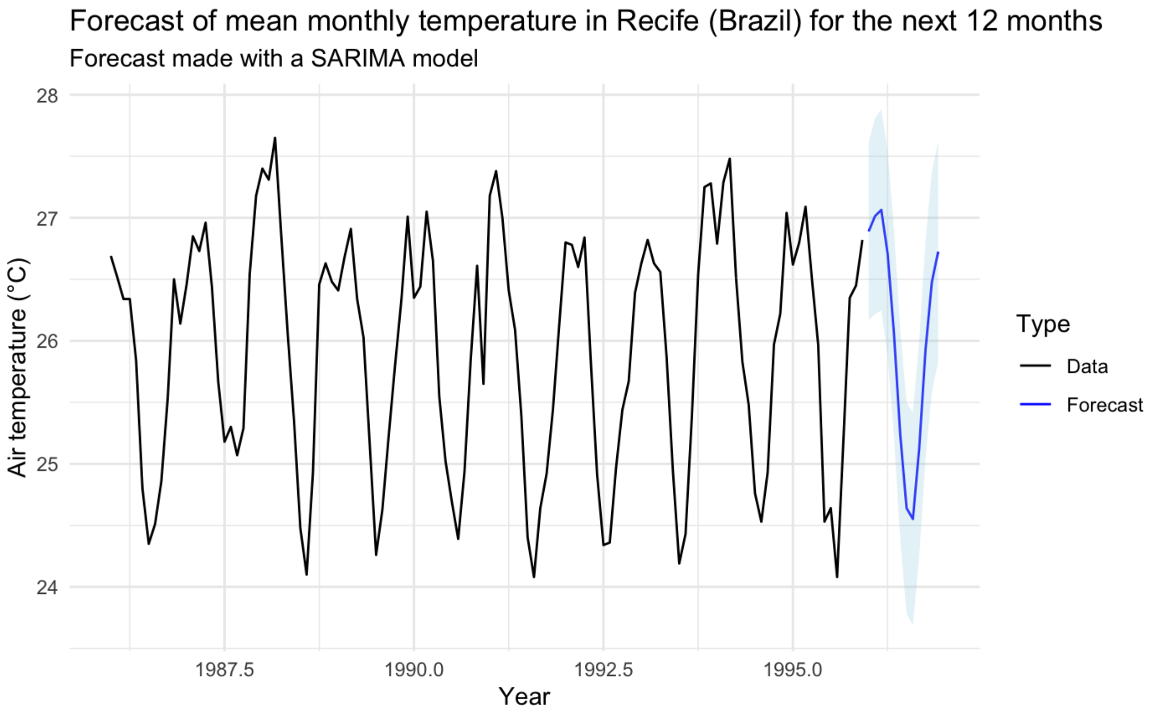


Figure 11 – *Forecast with a SARIMA model of the mean monthly air temperature in Recife (Brazil) for the next 12 months*

Then we would also like to compare this forecast with another one made via the Holt-Winters method.

Holt-Winters' seasonal method is a type of exponential smoothing forecasting method. These methods produce forecasts that consist in weighted averages of past observations where the weights decay exponentially as the observations get older.

Here, we use the Holt-Winters' additive method because our serie shows roughly constant seasonal variations (**??**). The equation is,

$$\hat{X}_{t+H} = l_t + hb_t + s_{t+h-m(k+1)} \tag{9}$$

where $m = 12$ is the number of seasons in a year, $k \equiv int(h-1)/m$ and l_t, b_t, and s_t are respectively the smoothing equations for the level l_t of the serie, the trend b_t and the seasonal component s_t.

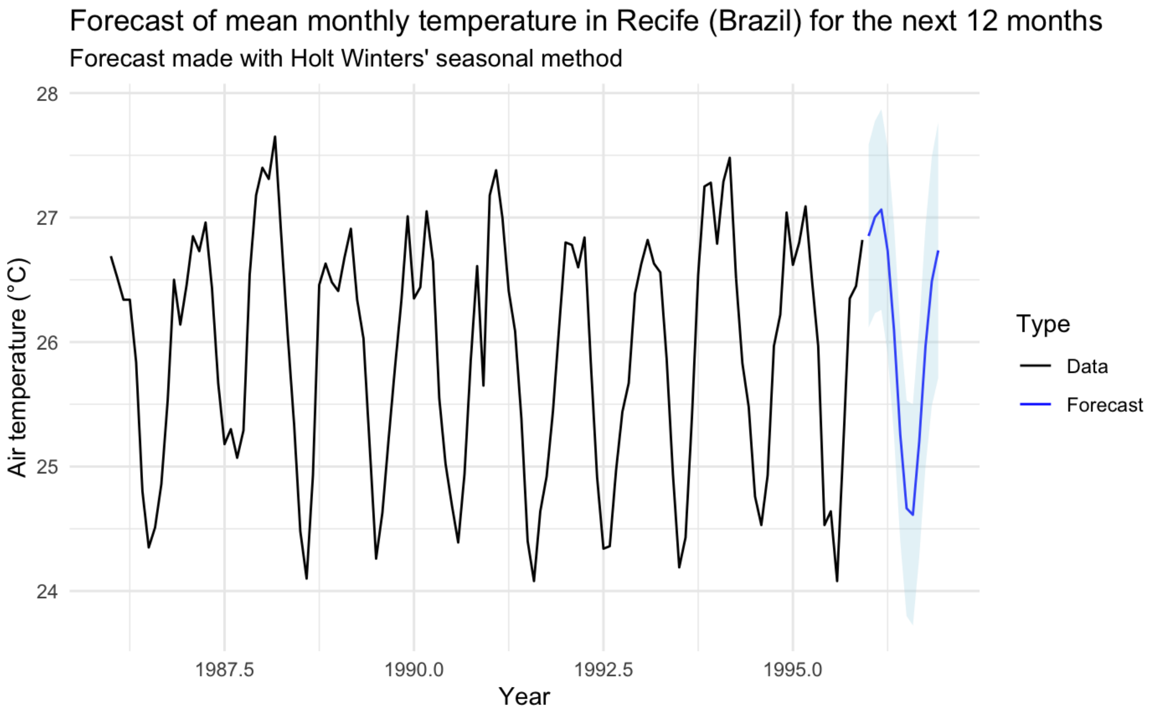


Figure 12 – *Forecast with a Holt-Winters' seasonal method of the mean monthly air temperature in Recife (Brazil) for the next 12 months*

5 Conclusion

We first plotted the time series and decomposed it into its trend, seasonal and remainder components assuming an additive decomposition. We saw this was an honest assumption because the seasonal variations were roughly constants over the years. Noticing the presence of seasons and trend, we first had to difference the time serie to remove them.

After doing so, we could analyse the ACF and PACF plots and made a first intuition about some possible SARIMA models we could pick in order to modelise our time serie. To confirm our intuition, we performed model selection and looked at different error criterions like the AIC and BIC. We selected the top 10% of models with the minimum AIC and decided to keep three models : two ... models () and one more complex model ().

Testing the coefficients significance, we removed the most complex model because it had one unsignificant coefficients for a level $\alpha = 5\%$. We also ensure that the residuals of the two other models were not serially correlated.

We finally compare the prediction power of these two lasting models comparing the last 2 years of the original data with a prediction made with these SARIMA models. We kept the one that minimized the mean sqaured error, that is the "model 3". As a last step, we predicted the mean monthly air temperature in Recife for the next 12 months using our chosen SARIMA model and also with an Holt-Winters seasonal method. We did not notice a clear difference between the two predictions. The two predicted ...

Appendix

Residual analysis

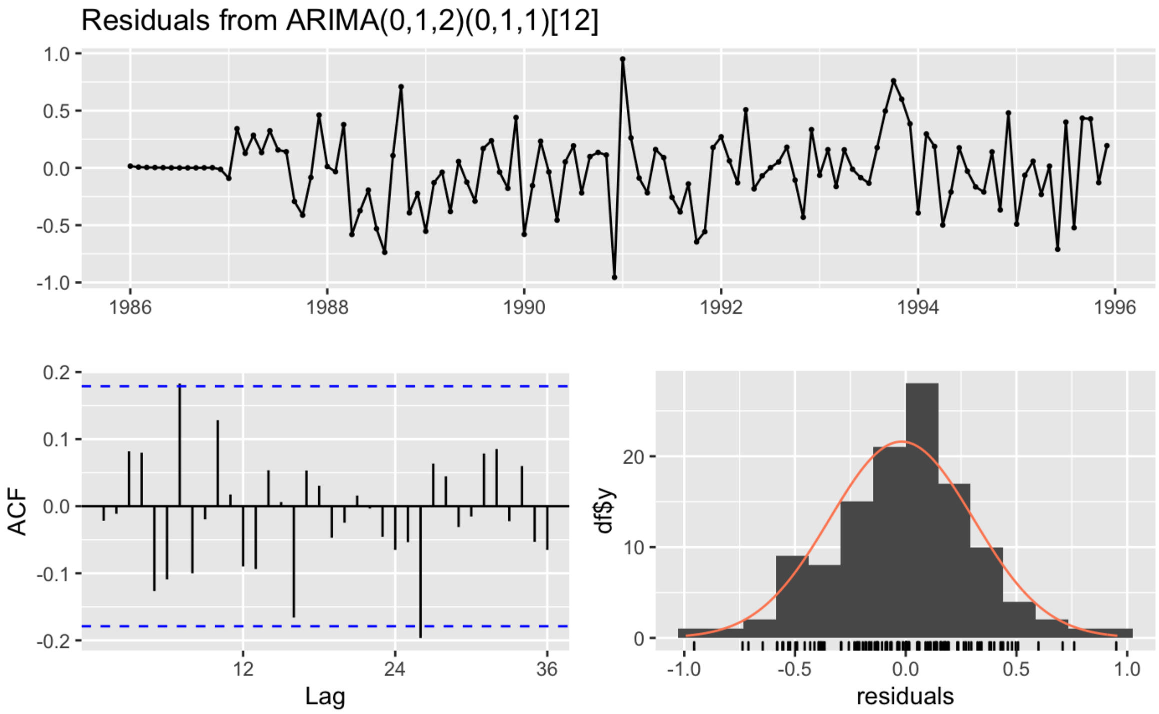


Figure 13 – *Residuals analysis for model 1.*

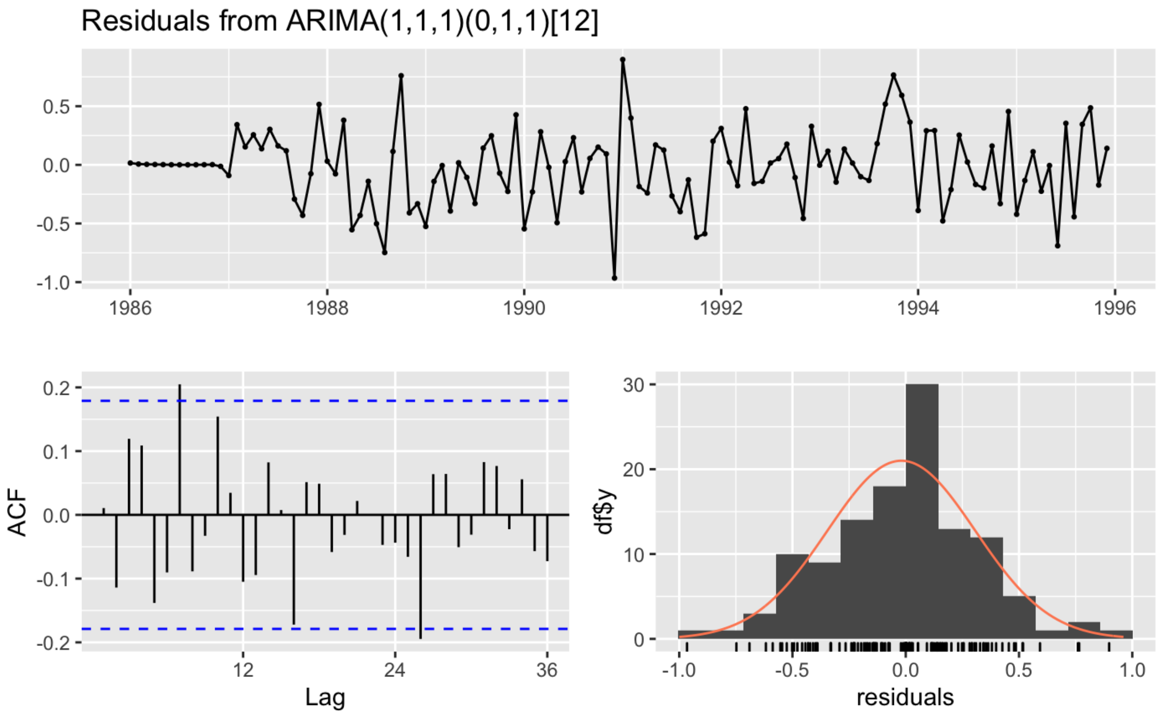


Figure 14 – *Residuals analysis for model 3.*